Experiments on Average Distance Measure

Vincenzo Della Mea[1], Gianluca Demartini[1],
Luca Di Gaspero[2], and Stefano Mizzaro[1]

[1] Dept. of Mathematics and Computer Science
[2] Dept. of Electrical, Management and Mechanical Engineering,
University of Udine, Udine, Italy
{dellamea, demartin, mizzaro}@dimi.uniud.it,
l.digaspero@uniud.it

Abstract. ADM (Average Distance Measure) is an IR effectiveness metric based on the assumptions of continuous relevance and retrieval. This paper presents some novel experimental results on two different test collections: TREC 8, re-assessed on 4-levels relevance judgments, and TREC 13 TeraByte collection. The results confirm that ADM correlation with standard measures is high, even when using less data, i.e., few documents.

1 Introduction

Common effectiveness measures for Information Retrieval Systems (IRSs) are based on the assumptions of binary relevance (either a document is relevant to a given query or not) and binary retrieval (either a document is retrieved or not). Several measures go beyond this and work with category relevance and ranked retrieval; almost no measures are available for the continuous relevance and retrieval case. One exception is ADM (*Average Distance Measure*) [1, 2, 3].

ADM measures the average distance between the amount of User Relevance Estimate (UREs, the actual relevances of documents) and the amount of System Relevance Estimates (SREs). ADM values lie in the [0, 1] range, with 0 representing the worst performance and 1 the performance of the ideal IRS. As discussed in detail in previous papers [1, 2, 3], ADM presents some nice theoretical properties; also, ADM has been experimentally validated on TREC and NTCIR data, with encouraging results, although the experimentation was somewhat limited. Indeed, an experimental confirmation of ADM effectiveness is both needed and difficult because very few data are available featuring continuous UREs and SREs, so that some approximations and assumptions are necessary.

The present work aims at providing further experimental evidence on the suitability of ADM for measuring the effectiveness of IRSs, especially when only a limited number of documents is available. In particular, this work aims at answering to the following two research questions: How many documents are needed to compute ADM in order to obtain results comparable to those of conventional measures like Average Mean Precision and R-Precision? What is the difference between computing ADM on the basis of two relevance levels or more?

In the experiments presented here, we used two document collections that include non-binary relevance scales (which are not continuous, yet provide more

M. Lalmas et al. (Eds.): ECIR 2006, LNCS 3936, pp. 492–495, 2006.

information than binary values): TREC 13 TeraByte, assessed on a 3-levels relevance scale, and TREC 8, re-assessed on a 4-levels relevance scale [4]. We compare ADM, by means of the Kendall's correlation, with the traditional effectiveness measures used by TREC, i.e., Mean Average Precision (MAP), R-Precision (R-Prec), and precision at N retrieved documents (P@N).

2 Experiments on TREC 13 TeraByte

The TREC 13 TeraByte test collection features data from 70 IRSs, 57 of which retrieved at least 1,000 documents for each topic. To study ADM effectiveness when considering only few documents, we compare the correlations among ADM@N (ADM calculated after N documents retrieved) and the reference measures.

For this test collection, Kendall's correlation between the two reference measures MAP and R-Prec is 0.82, whereas, as reported in Figure 1, the correlation

N	MAP	R-Prec
1000	0.90	0.91
500	0.90	0.92
400	0.89	0.92
300	0.90	0.93
200	0.92	0.94
100	0.90	0.90
90	0.88	0.88
80	0.87	0.86
70	0.87	0.87
60	0.88	0.87
50	0.87	0.85
40	0.87	0.85
30	0.86	0.84
20	0.82	0.81
10	0.76	0.76
5	0.73	0.72
1	0.57	0.59

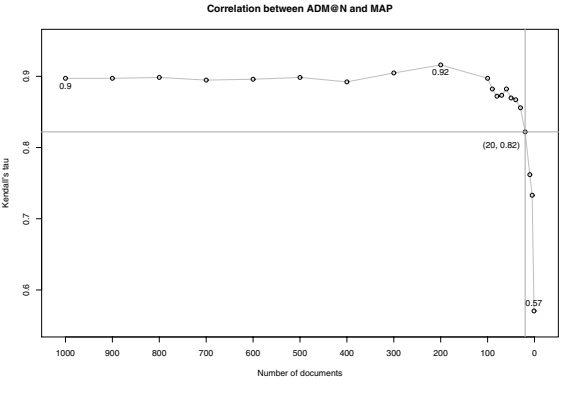

Fig. 1. Correlation between ADM@N and the two standard metrics MAP and R-Prec

		ADM							
	N	5	10	20	30	100	200	500	1000
Precision	5	**0.89**	0.88	0.88	0.86	0.82	0.76	0.75	0.75
	10	0.84	0.88	**0.91**	0.91	0.88	0.83	0.80	0.80
	20	0.82	0.85	0.92	**0.94**	0.89	0.85	0.82	0.82
	30	0.81	0.83	0.91	**0.94**	0.87	0.81	0.78	0.77
	100	0.72	0.74	0.81	0.85	**0.94**	0.93	0.90	0.90
	200	0.71	0.75	0.79	0.82	0.91	**0.98**	0.94	0.93
	500	0.67	0.71	0.75	0.77	0.85	0.92	**0.99**	0.97
	1000	0.66	0.68	0.72	0.75	0.83	0.87	**0.92**	**0.92**

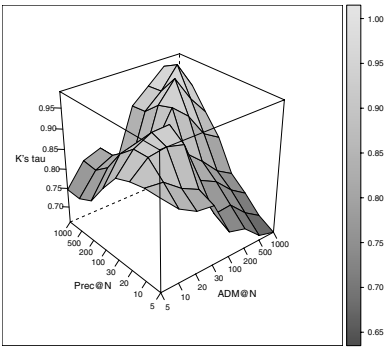

Fig. 2. Correlation between ADM@N and P@N

between ADM@N and the reference measures is higher than 0.82 for $N \geq 20$. This suggests that, at least for this test collection, ADM calculated on only the first 20 documents provides the same information value as R-Prec and MAP computed on the whole set of retrieved documents.

Figure 2 shows how the correlation between ADM@N and P@N varies depending on the value of N. As expected, the higher correlation values for each N (shown in boldface) lie on the diagonal of the table or in its proximity, so that the two measures correlates most for equal (or very close) N values. This confirms that ADM@N and P@N measure similar phenomena.

3 Experiments on 4-Levels Relevance TREC

Sormunen [4] has re-assessed 18 topics from TREC 7 and TREC 8 using 4 levels of relevance (0, 1, 2 and 3). For the sake of applying traditional binary measures, these levels can (and have to) be binarized as either a *Rigid* mapping (levels 0 and 1 become 0, levels 2, and 3 become 1) or a *Relaxed* mapping (level 0 becomes 0, levels 1, 2, and 3 become 1).

We calculated ADM using both the 4 levels of relevance (denoted by ADM[4]) and the rigid and relaxed binary data (ADM[2rig] and ADM[2rel], respectively). These three ADM values were then compared with the reference measures MAP and R-Prec calculated on the Sormunen data (see Table 1). We then compared Sormunen and ADM values with the original MAP and R-Prec measures calculated on the TREC 8 data (see Table 2 and Figure 3).

ADM computed on the binary relaxed mapping has a higher correlation with the reference measures than ADM computed on the 4 levels of relevance. We conjecture that this phenomenon is related to the TREC evaluation rules: the TREC guidelines state that a document is judged relevant if any piece of it is relevant, thus the relaxed mapping matches better with the reference measures calculated by the original TREC assessments. This is a confirmation of the results

Table 1. Correlation between ADM and R-Prec and MAP. All measures are computed on the basis of the Sormunen's 4 levels reassessment.

	ADM[2rig]	ADM[2rel]	ADM[4]
R-Prec[rig]	0.75	0.70	0.77
R-Prec[rel]	0.80	0.83	0.90
MAP[rig]	0.41	0.40	0.39
MAP[rel]	0.67	0.67	0.64

Table 2. Correlation between ADM computed on the basis of the Sormunen's 4 levels reassessment and the original TREC 8 measures

		Sormunen						
		ADM[2rig]	ADM[2rel]	ADM[4]	R-Prec[rig]	MAP[rig]	R-Prec[rel]	MAP[rel]
TREC 8	ADM	0.80	0.94	0.86	0.69	0.39	0.82	0.66
	MAP	0.79	0.85	0.82	0.72	0.43	0.82	0.79
	R-Prec	0.79	0.84	0.80	0.68	0.46	0.78	0.79

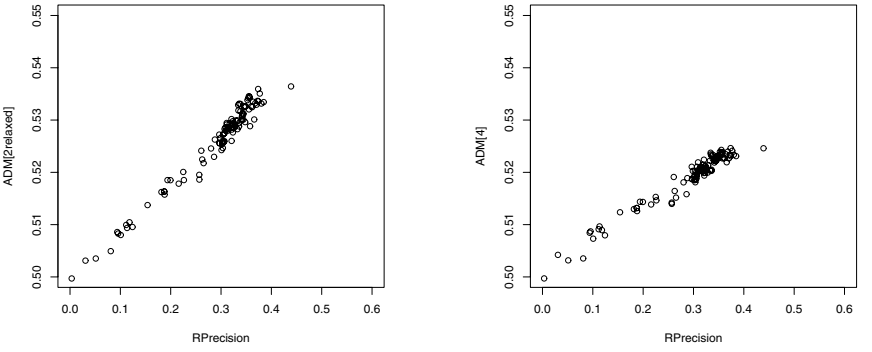

Fig. 3. Correlation between ADM[2rel] and R-Prec and between ADM[4] and R-Prec

shown in [4]. However, differences between ADM calculated on rigid and relaxed data are lower than those between either MAP or R-Prec calculated on rigid and relaxed data. This fact may be interpreted as either a greater robustness of ADM or a lower sensitivity to relevance variations, and thus needs further experimentations to be fully understood.

4 Conclusions and Future Work

The results on ADM presented in this paper are to be considered still preliminary. However, when considered together those already presented in [1, 2, 3], give insights on the capabilities of ADM as an effectiveness measure for information retrieval systems. In particular, the results show that ADM correlation with standard measures (R-Prec, MAP, P@N) is high, and that the correlation is still high also when using just few documents. The latter capability makes ADM easier to use for IRS evaluation than traditional binary measures.

In the future, we plan to further study the phenomena emphasized above; we are experimenting with ADM on INEX 2004 data and we intend to build an IRS capable of estimating the amount of relevance on a continuous scale.

References

1. V. Della Mea, L. Di Gaspero, and S. Mizzaro. Evaluating ADM on a four-level relevance scale document set from NTCIR. In *Proceedings of NTCIR Workshop 4 Meeting - Supplement Vol. 2*, pages 30–38, 2004.
2. V. Della Mea and S. Mizzaro. Measuring retrieval effectiveness: A new proposal and a first experimental validation. *JASIS&T*, 55(6):530–543, 2004.
3. S. Mizzaro. A new measure of retrieval effectiveness (Or: What's wrong with precision and recall). In T. Ojala, editor, *International Workshop on Information Retrieval (IR'2001)*, pages 43–52, 2001.
4. E. Sormunen. Liberal relevance criteria of TREC - Counting on negligible documents? In K. Jarvelin, M. Beaulieu, R. Baeza-Yates, and S. Myaeng, editors, *Proceedings of the 25th ACM SIGIR Conference*, pages 324–330, August 11–15 2002.